

# 群体移动数据：模型、查询与应用

## 1、引言

群体移动广泛存在于人类社会和生物世界，如城市中人群如潮汐般白天涌入城市中心而傍晚则向周边居民区分散，候鸟秋天集体飞往南方越冬而春天返回北方。群体移动模式的研究对于了解群体行为模式和规律具有十分重要的意义。

现有的群体移动数据分析主要基于移动数据模型。移动数据模型在时空数据的基础结构上，提供了诸多时空操作，如时空区间投影、时空区间交叉等。移动数据模型主要以时空为研究对象，计算时空区间内存在的移动物体，可以表示为 $\tau(\alpha) = time \rightarrow \alpha$ ，其中 $\alpha \in \{point, line, region\}$ 。在群体移动数据分析中，很多应用是从群体而不是单个对象出发研究其运动模式。另外，群体移动数据分析中群体对象比时空属性更受关注。因此，以群体对象为基础的数据模型对于分析群体移动模式更为直接。

群体移动数据规模较大，查询和分析需要消耗较多的资源，MR模型提出了简单有效的方案。然而可视分析技术让分析人员能够通过图形界面交互式探索数据，因此要满足数据分析要求仍面临两方面的挑战。

- 适合群体移动分析的数据组织。尽管MR模型具有很强的适应性，但是对于特殊的数据仍需要定制数据组织方式。（举例列存储，PAX）对于群体移动数据目前仍没有能够明显提高分析性能的数据组织方式。
- 实时查询。要做到交互式分析，查询响应时间要短。面对大数据量的群体移动数据，不仅要求查询算法复杂度低，而且数据能够快速索引。与数据库系统不同HDFS上的数据通常并不包含索引，因此每一次计算都需要遍历所有数据。Hadoop++，HAIL等提出了一些通用解决方案自适应地为数据建立索引。不同的是，群体移动数据包含时空属性，很难用常规的索引方法表示。

本文提出群体移动数据模型，模型以群体为主要研究对象，抽象群体移动查询操作，为多种群体移动数据分析提供服务。主要贡献包含以下三点：

- 群体移动数据模型。
- HDFS上的移动数据索引结构。
- 面向可视分析的查询算法

## 2、相关工作

### 2.1 移动数据模型

### 2.2 基于MR模型的时空查询

### 2.3 群体移动数据的可视分析

## 3、群体移动模型

### 3.1 模型

### 3.2 ping-pang effect 处理

### 3.3 查询模式

- 群体运动轨迹查询：查询一类群体特定时间范围内经过的空间区域
- 群体迁移查询：查询某时间段内从一个空间区域移动到另外一个空间区域的群体

- 范围查询：指定时间范围和空间区域查询包含的移动群体

#### 4、索引结构

基本想法是能在 HDFS 上实现全局的、自适应的索引。用户上传数据后自动建立索引，而不需要额外操作。另外，索引支持动态增长，也就是说有新数据加入时，自动添加索引到原来的索引上。这里的难度在于：如何自动建立全局索引。（目前正在实现）

#### 5、查询算法

计划使用 spark 提供的计算模型接口实现，主要原因来自两方面：一是 spark 的 MR 计算中间结果保存于内存，要比 Hadoop 计算性能更高（Hadoop 中间结果存于磁盘）；二是 spark 计算模型编程更容易。

#### 6、应用

此部分列举利用交互查询得到的可视分析的结果。

#### 7、性能分析